IBIS2010

補助情報を用いた情報源符号化

松嶋敏泰
早稲田大学

統計的学習理論と情報理論

（関係するものですぐ思いつくのは）

情報源符号化（冗長度を削減）

ユニバーサル情報源符号

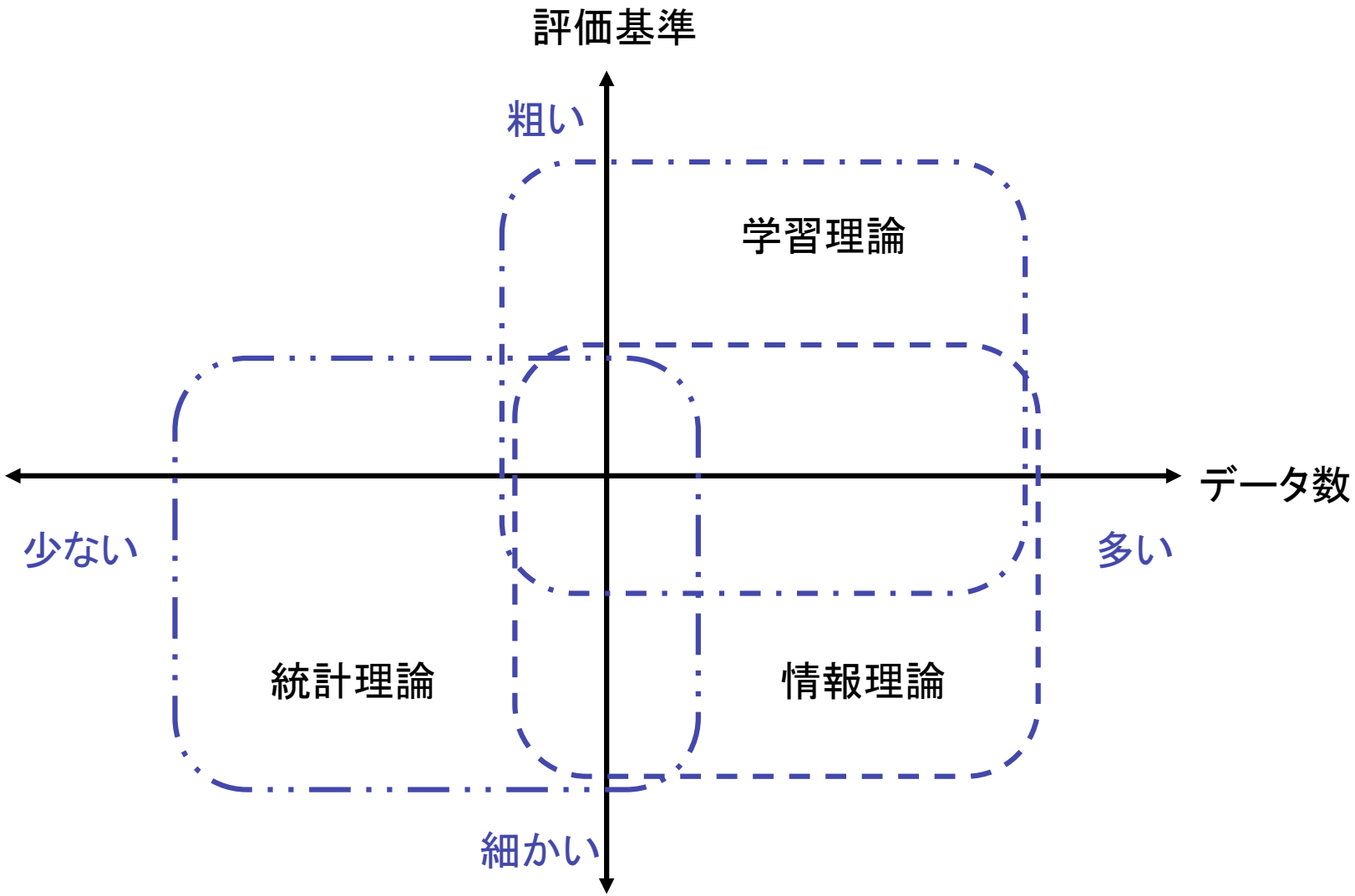
予測，統計的複雑度，
モデル選択，．．．

通信路符号化（冗長度を付加）

復号法（誤り訂正符号，CDMA，MIMO）

グラフィカルモデル，
事後確率計算，BP，
和積アルゴリズム，平
均場近似，．．．

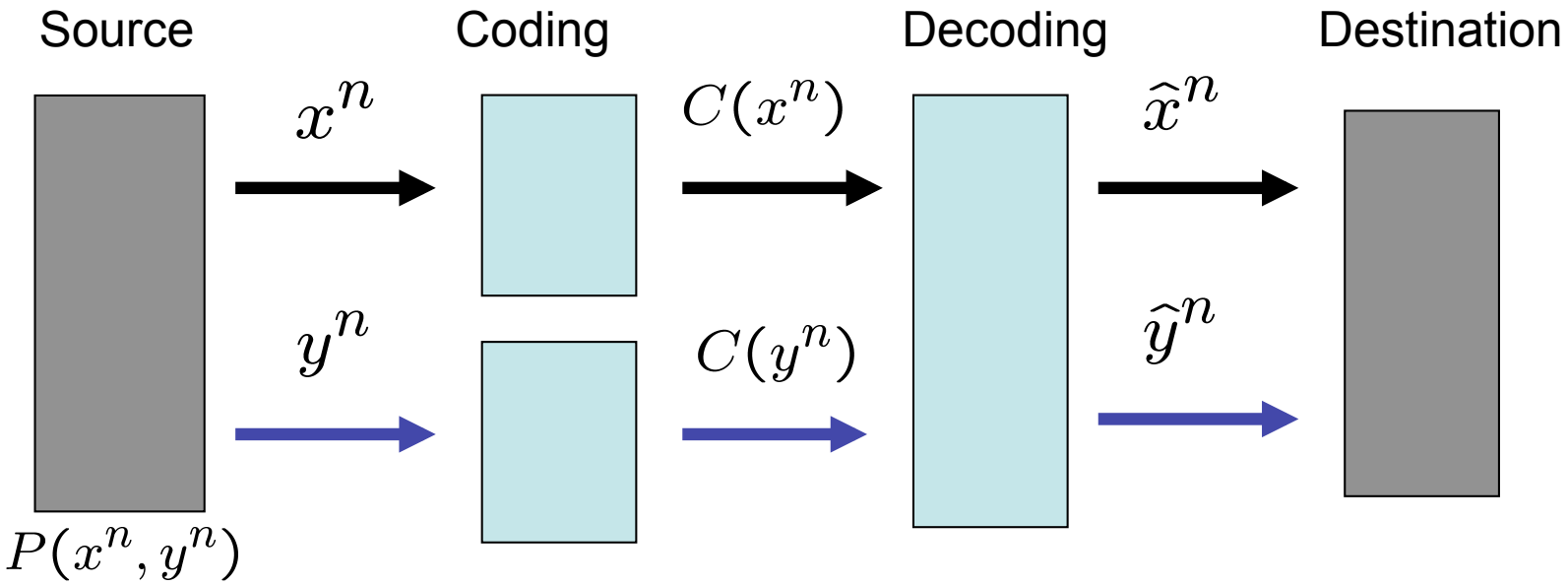
学習理論，統計理論，情報理論
（無理して特徴付けると）



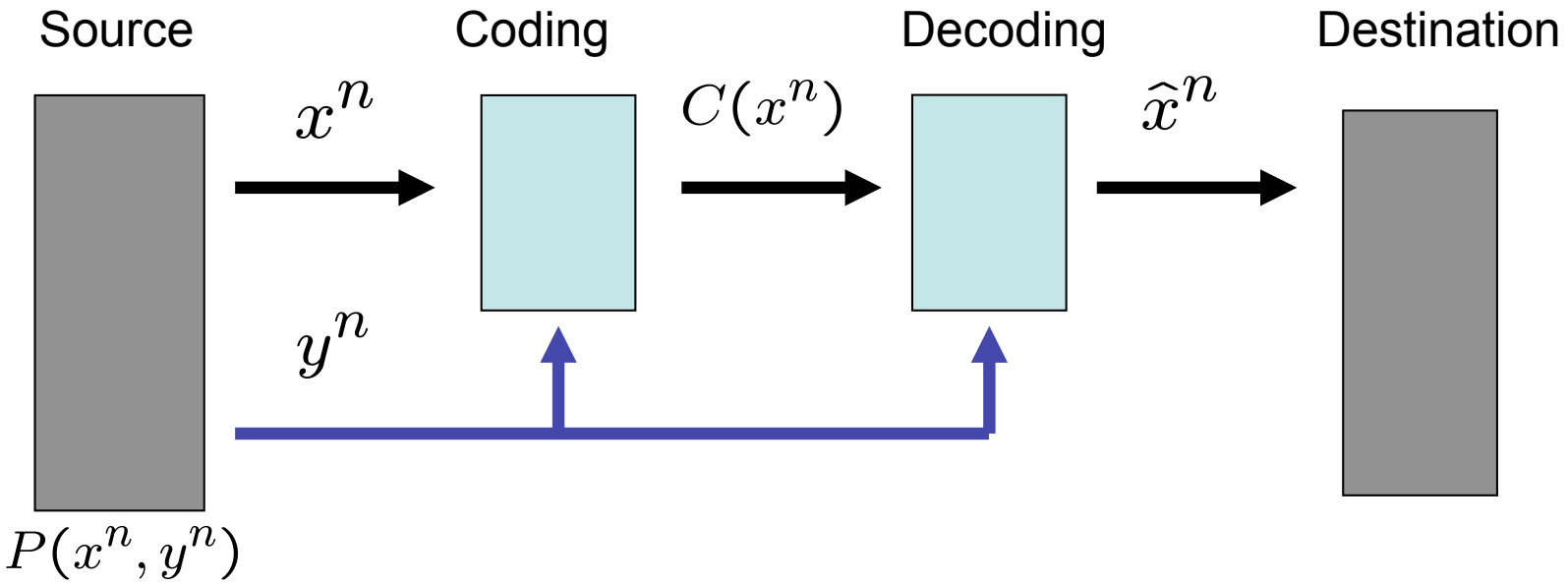
情報理論の特徴

- 評価基準が明確（復号誤り率，レート）
- 評価基準の限界をまず示す
- その限界を達成する符号化，復号化を求める
- 実用的アルゴリズムを求める（計算量，メモリー量）

0. Slepian –Wolf Problem



1.Surece coding with side information

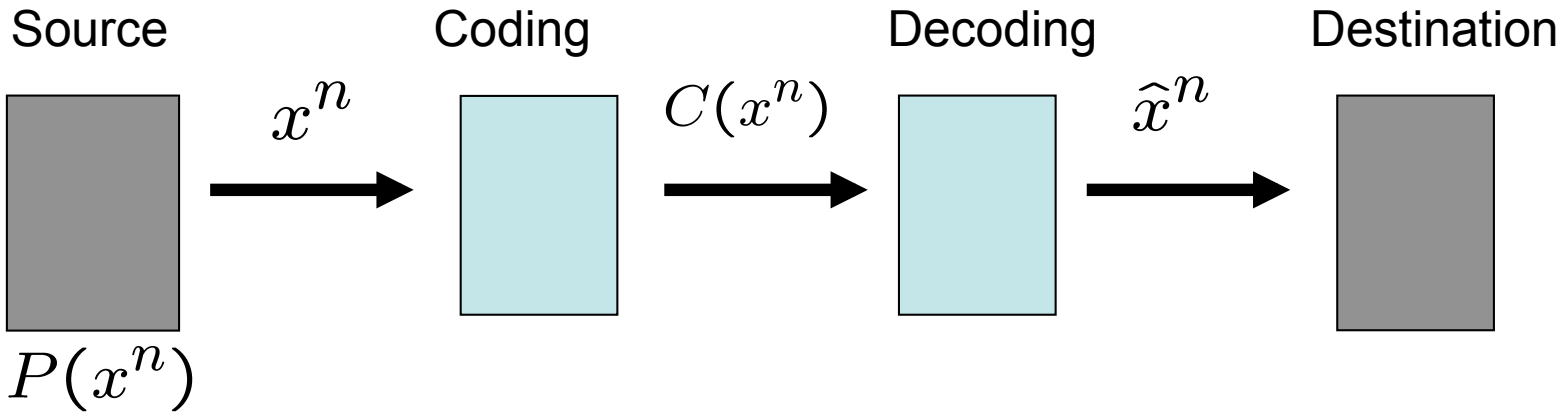


Distortion less: $x^n = \widehat{x}^n$

Mean code length: $\lim_{n\to\infty} 1/nE[|C(X^n)|] \geq H(X|Y)$

2. Source coding

$$x^n = x_1, x_2, \cdots, x_n (x_i \in A)$$



Distortion less: $x^n = \hat{x}^n$

Mean code length: $\lim_{n\to\infty} 1/nE[|C(X^n)|] \geq H(X)$

Ideal code length: $|C(x^n)| = -\log P(x^n)$

If the distribution of a source is known, we can construct the optimal source code.

eg) Arithmetic codes

ユニバーサル情報源符号化と学習理論

最適な符号化

ブロック符号　$C(x^n)$　　$|C(x^n)| = -\log P(x^n)$

逐次符号　$C(x_t|x^{t-1})$　　$|C(x_t|x^{t-1})| = -\log P(x_t|x^{t-1})$

ユニバーサル符号問題　$P(x_t|x^{t-1})$　が未知

損失関数:　$\log \hat{P}(x_t|x^{t-1}) - \log P(x_t|x^{t-1})$

累積損失:　符号長

累積リスク:　期待符号長

補助情報つき情報源符号化と学習理論

$x^n : x_1, x_2, \cdots, x_n$ 目的変数

$y^n : y_1, y_2, \cdots, y_n$ 説明変数 （補助情報）

$x_1, x_2, \cdots, x_{t-1}, x_t$

$y_1, y_2, \cdots, y_{t-1}, y_t$

回帰問題，分類問題

$$\widehat{P}(x_t | y_t, x^{t-1}, y^{t-1})$$

補助情報つき情報源符号化の問題

情報理論的にはほとんど解決

補助情報つき情報源符号化についてはZiv[1984], Subarahmanya[1995], Muramatsu[1998], Yan[2001],Uematsu[2002]等の研究があり，i.i.d. や定常エルゴード情報源等について，漸近最良の符号$\frac{C \log n}{n}$で平均符号長が条件付エントロピーに近づくことは示されている.

学習理論の問題を考えると，リスク関数を最小化するダイレクトで詳細な性能解析が望まれる（最速でエントロピーに近づく）.

ベイズ符号で補助情報つき情報源符号化を考える

3. Why Bayes codes?

The advantage of Bayes codes?

1) Precise evaluation of the code length:

We can evaluate the exact asymptotic code length of Bayes codes as far as constant terms

2) The highest convergence rate:

The convergence rate of Bayes codes achieves the highest order for universal codes

3) Maximin codes and Minimax codes:

The class of Bayes codes includes Maximin codes (Jeffrys' prior) and Minimax codes

4-1. Bayes codes

Known ← a parametric distribution: $P(x^n|\theta)$

Unknown ← k-dimensional real parameter vector: $\theta \in \Theta \subset \Re^k$

Under the condition that the **Arithmetic Coding** is used for coding, the decision of **encoding prob** $\hat{P}(x^n)$ is the main problem in universal coding.

The optimal solution from the view point of Bayes strategy

Block code: $$\hat{P}_B(x^n) = P(x^n) = \int_\Theta P(x^n|\theta)dP(\theta),$$

A block sequence x^n is encoded to a code word $C(x^n)$ simultaneously

Predictive code: $$\hat{P}_P(x_t|x^{t-1}) = P(x_t|x^{t-1}) = \int_\Theta P(x_t|x^{t-1}),\theta))dP(\theta|x^{t-1})$$

A symbol x_t given x^{t-1} is encoded to a code word in order

4-2. the code length of Bayes codes

The code length of the Bayes block code is the same as that of the Bayes predictive code.

$$-\log \hat{P}_B P(x^n) = \sum_{t=1}^{n} -\log \hat{P}_P(x_t|x^{t-1})$$

Precise evaluation [Clarck&Barron90]:

$$E_\theta[-\log P(X^n)] = H(X^n|\theta) + \frac{k}{2}\log\frac{n}{2\pi e} + \log\frac{\sqrt{\det I_X(\theta)}}{p(\theta)} + o(1).$$

Fisher information

This convergence rate achieves the highest order for universal codes

5. Precise evaluation of the code length of the Bayes codes with side information

The average code length of Bayes codes with side information under some assumption is given by

$$
\begin{aligned}
& -E_{X^nY^n}\left[\log P_c(x^n|y^n)\right] \\
&= -E_{X^nY^n}\left[\log P(x^n|y^n,\nu^*)\right] + \frac{k}{2}\log\frac{n}{2\pi e} \\
&\quad + \frac{1}{2}\log\det I_{X|Y}(\nu^*) + \log\frac{1}{p(\nu^*)} + o(1),
\end{aligned}
\tag{1}
$$

where $f(\nu)$ is the prior probability of ν.

Conditional Fisher information:

$$
I_{X|Y}(\nu) = -\lim_{n\to\infty}\frac{1}{n}E_{X^nY^n}\left[\frac{\partial^2 \log P(X^n|Y^n,\nu)}{\partial\nu(\partial\nu)^T}\right]
$$

$$
E_\theta[-\log P(X^n)] = H(X^n|\theta) + \frac{k}{2}\log\frac{n}{2\pi e} + \log\frac{\sqrt{\det I_X(\theta)}}{p(\theta)} + o(1).
$$

6. Bayes codes (unknown model)

Known ← a parametric distribution class: $P(x^n|\theta(m), m)$

Unknown ← k-dimensional real parameter vector: $\theta \in \Theta \subset \Re^k$

Unknown ← parametric model: $m \in M$

The encoding probability of Bayes codes:

$$\hat{P}_B(x^n) = \sum_{m \in M} \int P(x^n|\theta_m, m) p(\theta_m|m) P(m) d\theta.$$

Expectation is taken all over the model class

Precise evaluation for the code length for unknown model problem is given by[Gotoh & Matsushima98]

The disadvantage of Bayes codes?

The complexity of the coding calculation

Assuming the conjugate prior as the prior probability of θ, the integral does not need for calculating the coding probability.

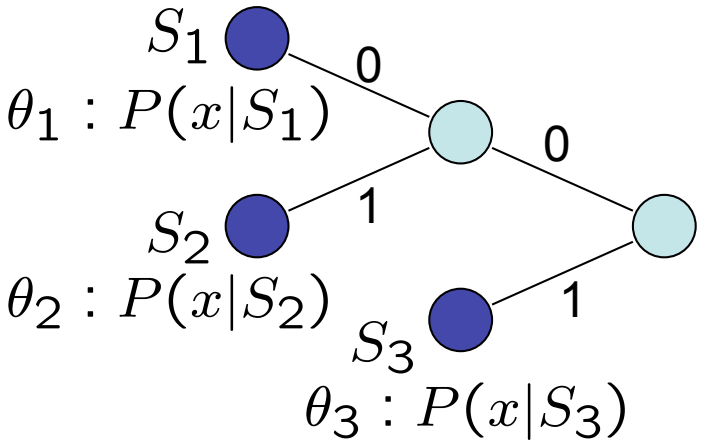
Integral is taken all over the parameter space

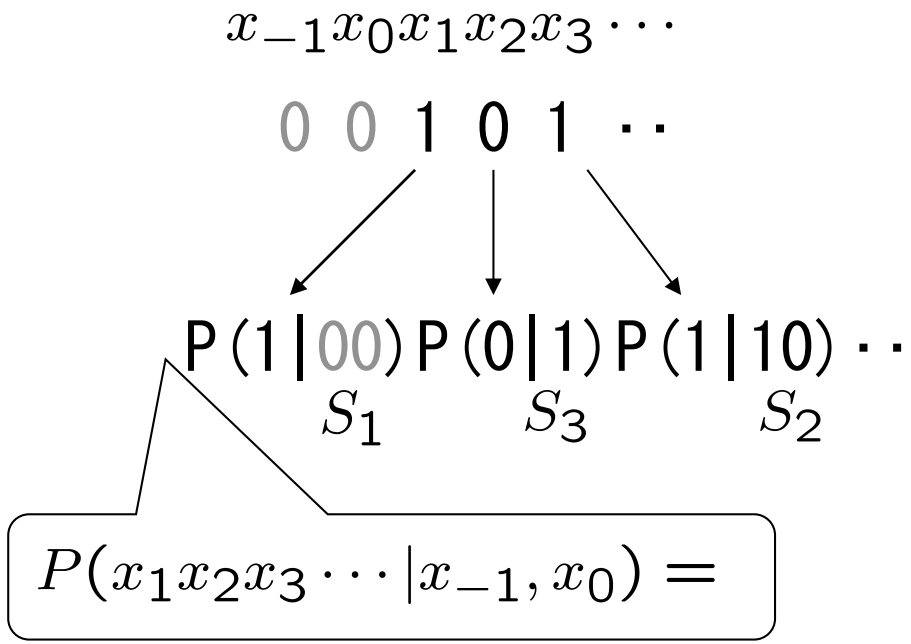
$$\hat{P}_B(x^n) = \sum_{m \in M} \int_{\Theta} P(x^n|\theta_m, m) p(\theta_m|m) P(m) d\theta.$$

Summation is taken all over the model class

Assuming a special prior distribution on the models, some efficient algorithms can be constructed.

7-1. Context tree models





A context tree(FSMX) model *m* is represented by a context tree(or a set of contexts) and a set of the parameters on the leaves :

$$P(x^n|m,\theta(m))$$

$\{S_1, S_2, S_3\}$ $\{\theta_1, \theta_2, \theta_3\}$

8. Bayes codes for Context tree models(1)

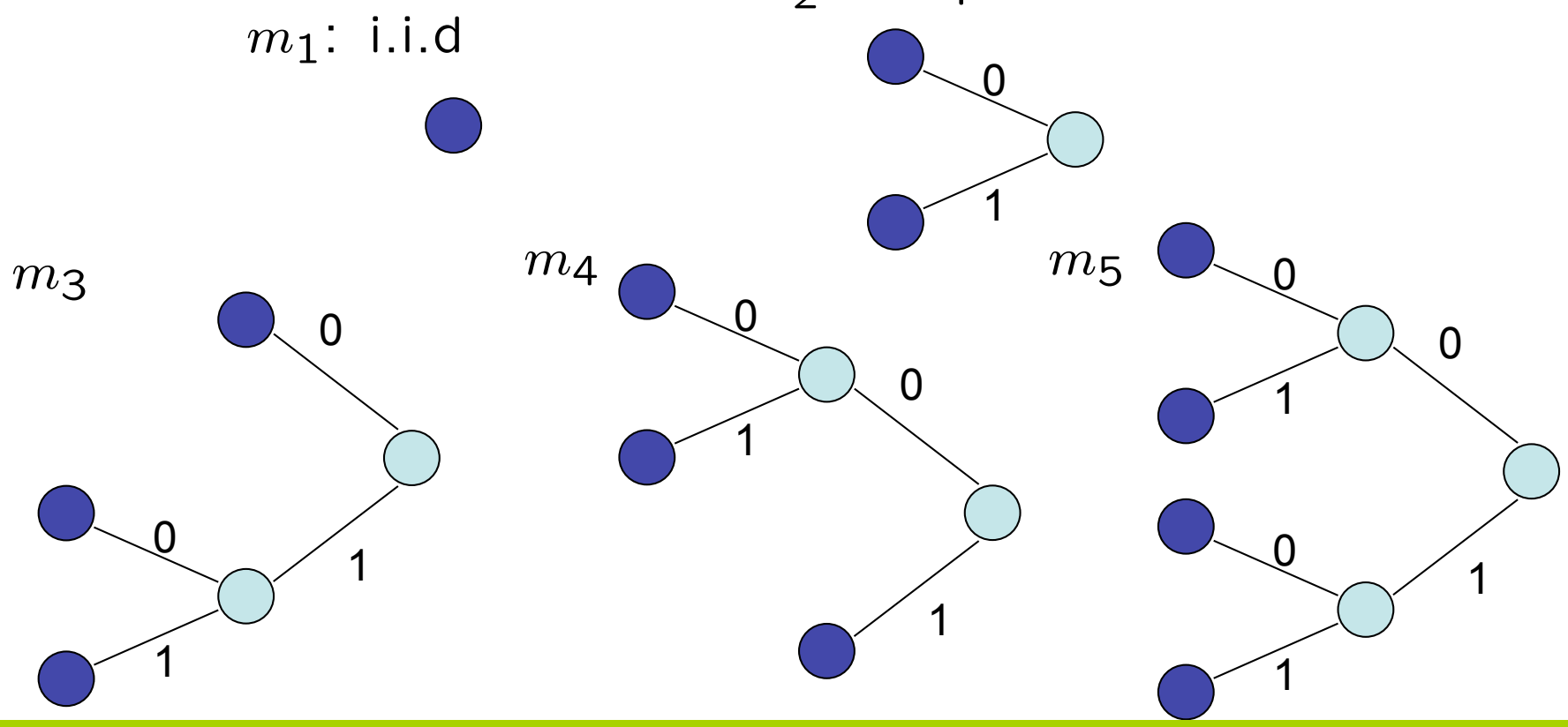
Unknown model $m_i = \{S_1, S_2, S_3 \cdots\} \in M$

Unknown parameters $\theta(m) = \{\theta_1, \theta_2, \theta_3 \cdots\} \in \Theta(m)$

M: the class of context trees up to depth 2

m_1: i.i.d

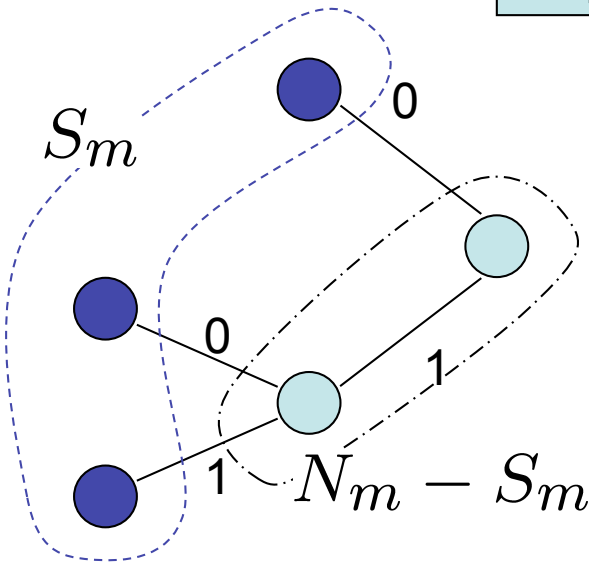
m_2: simple Markov model



For calculation of the Bayes code, the expectation is taken over all context tree models whose number is $O(2^{|X|^{d-1}})$

8. Bayes codes for Context tree models(2)

A special class of the prior distribution of *m*



We assume that the prior probability $P(m)$ is represented by the following formula:

$$P(m) = \prod_{s \in N_m - S_m} (1 - q(s)) \prod_{s_L \in S_m} q(s_L). \quad (1)$$

where

$$q(s) = \frac{P(S_{m_s})}{\sum_{S \in S^d} P(S_{m_s} \cup S)}, \quad (2)$$

where $m_s \in \{m | s \in S_m\}$, S_d denotes the class of the set of descendant nodes of s that construct a complete tree.

The prior probability of m is represented by $\{q(s)|s \in N_m\}$. Moreover, the posterior probability $P(m)$ is also represented by $\{q(s|x^t)|s \in N_m\}$.

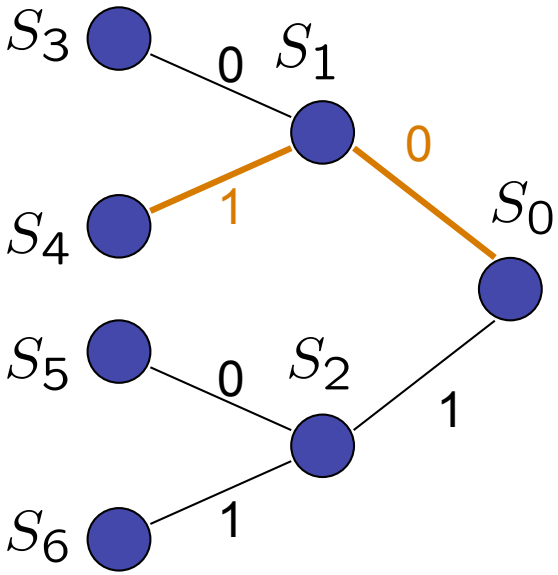
The CTW algorithm is regarded as the Bayes code assuming *q(s) = 1/2*

8. Bayes codes for Context tree models(3)

The efficient coding algorithm is given by using the tree to which all context tree models in *M* are merged and the special prior.

$\hat{P}(x^n)$ ⇐ Block codes: CTW algorithm

$\hat{P}(x_{t+1}|x^t)$ ⇐ Predictive codes: Predictive Bayes algorithm for Context tree models [Matsushima95]



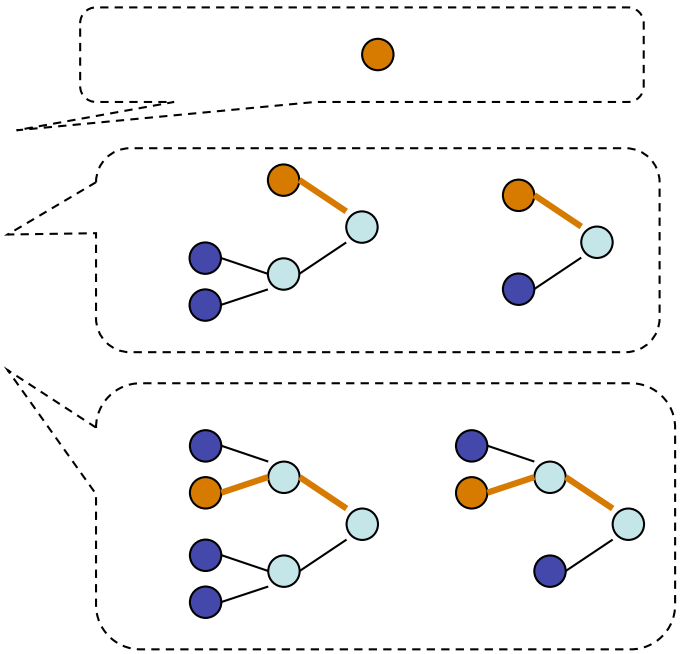
$\cdots x_{t-1} x_t x_{t+1}$

$\cdots 1 \quad 0 \quad 0$

Posterior probability of models

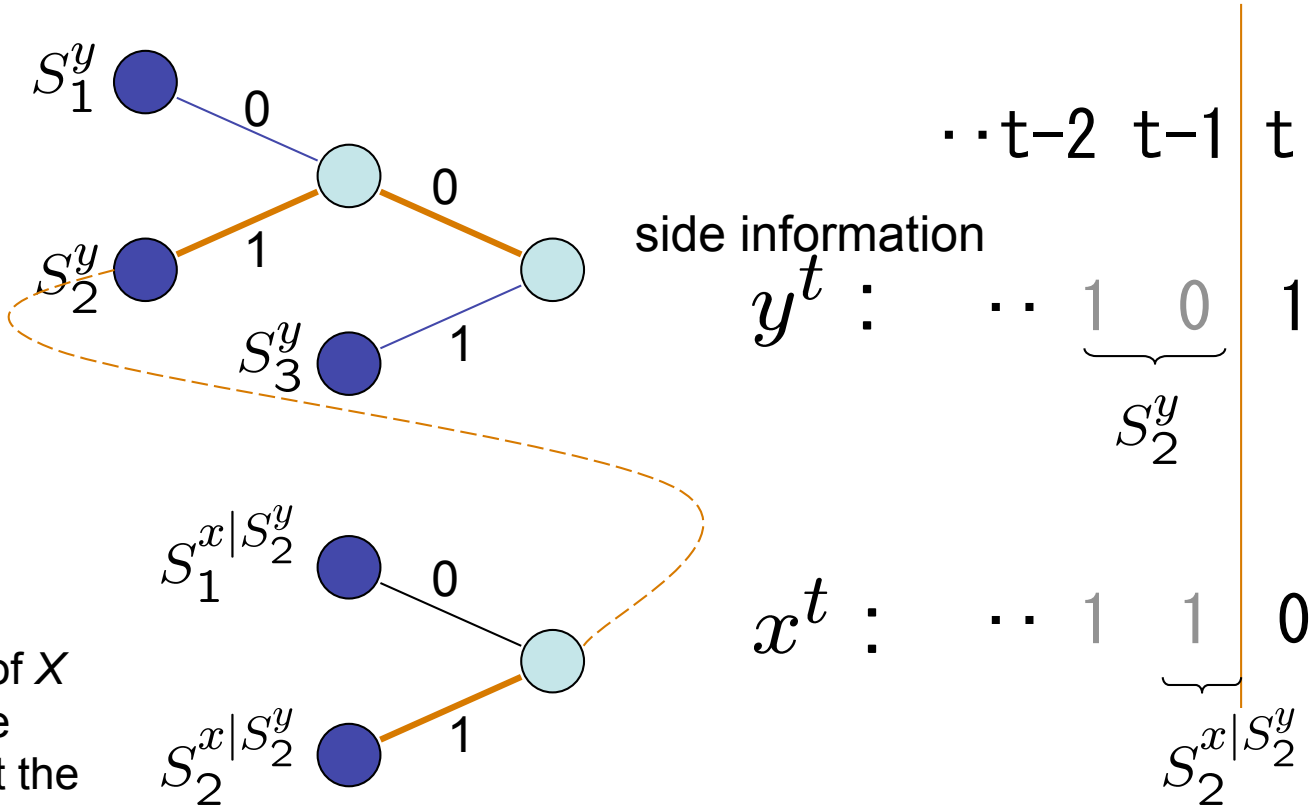
$$P(0|10) = \hat{P}(0|S_0)\hat{q}(S_0|x^t) + \hat{P}(0|S_1)\hat{q}(S_1|x^t) + \hat{P}(0|S_4)\hat{q}(S_4|x^t)$$

$$\hat{P}(0|S_i) = \frac{n(0|S_i) + \alpha/2}{n(S_i) + \alpha}$$



By using *q(s)*, the complexity is deduced to *O(d)*.

9. Side information Context tree models



The context of *X* is *1* under the condition that the context of *Y* is *10*

$$P(x_t = 0|x^{t-1} = \cdots 11, y^{t-1} = \cdots 10) = P(x_t = 0|S_2^{x|S_2^y}, S_2^y)$$

A side information context tree model is represented by

$$P(x_t|x^{t-1}, y^{t-1}) = P(x_t|S^{x|S^y}, \nu)P(S^y|\mu)$$

9. A special prior on Side-information context tree models

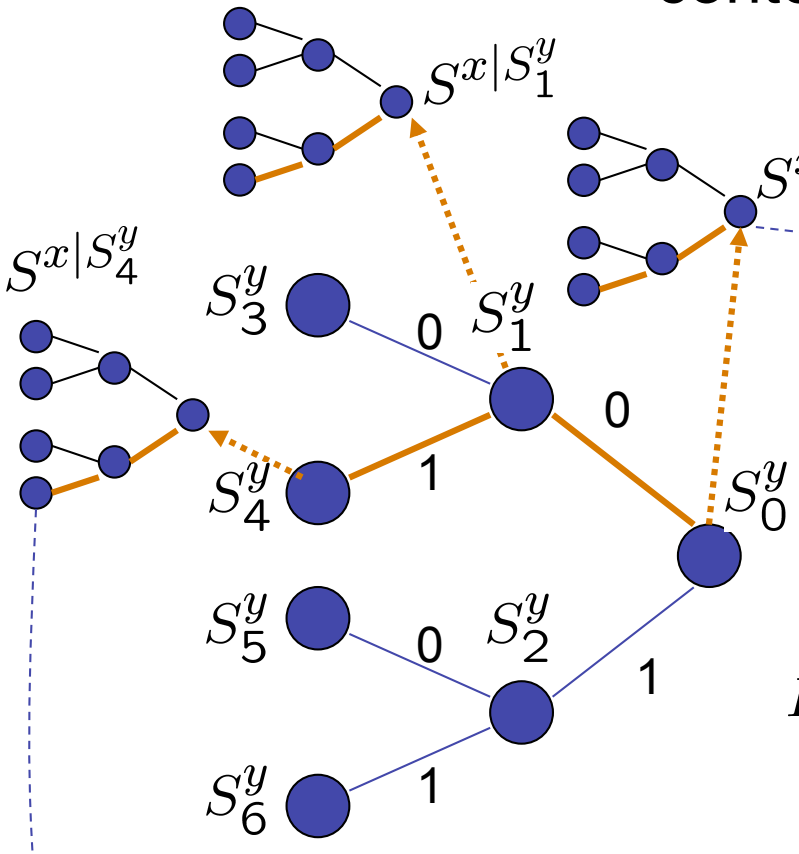
A special class of the prior distribution of *m*

[Matsushima 05]

We assume that the prior probability $P(m)$ of a side information context tree model is represented by the following formula:

$$P(m) = \prod_{s^y \in N_m^y - S_m^y} (1 - q(s^y)) \prod_{s_L \in S_m^y} q(s_L^y) \prod_{s^x \in N_m^{x|s_L^y} - S_m^{x|s_L^y}} (1 - q(s^x|s_L^y)) \prod_{s_L^x \in S_m^{x|s_L^y}} q(s_L^x|s_L^y). \quad (1)$$

10. Algorithm of Bayes codes for Side-information context tree models



Posterior probability of states of X

Posterior probability of states of Y

$$
\begin{aligned}
P(x_{t+1}=0|x^t=11,y^t=10) \\
&= \hat{P}(0|S_0^{x|S_0^y})\hat{q}(S_0^{x|S_0^y}|x^t,y^t)\hat{q}(S_0^y|x^t,y^t) \\
&+ \hat{P}(0|S_2^{x|S_0^y})\hat{q}(S_2^{x|S_0^y}|x^t,y^t)\hat{q}(S_0^y|x^t,y^t) \\
&\quad \cdots \\
&+ \hat{P}(0|S_6^{x|S_4^y})\hat{q}(S_6^{x|S_4^y}|x^t,y^t)\hat{q}(S_4^y|x^t,y^t)
\end{aligned}
\tag{1}
$$

11. Precise evaluation of the code length of the Bayes codes

The average code length of Bayes codes with side information under some assumption is given by

$$
\begin{aligned}
& -E_{X^nY^n}\left[\log P_c(x^n|y^n)\right] \\
&= -E_{X^nY^n}\left[\log P(x^n|y^n,\nu^*)\right] + \frac{k}{2}\log\frac{n}{2\pi e} \\
&\quad + \frac{1}{2}\log\det I_{X|Y}(\nu^*) + \log\frac{1}{p(\nu^*)} + o(1),
\end{aligned}
\tag{1}
$$

where $f(\nu)$ is the prior probability of ν.

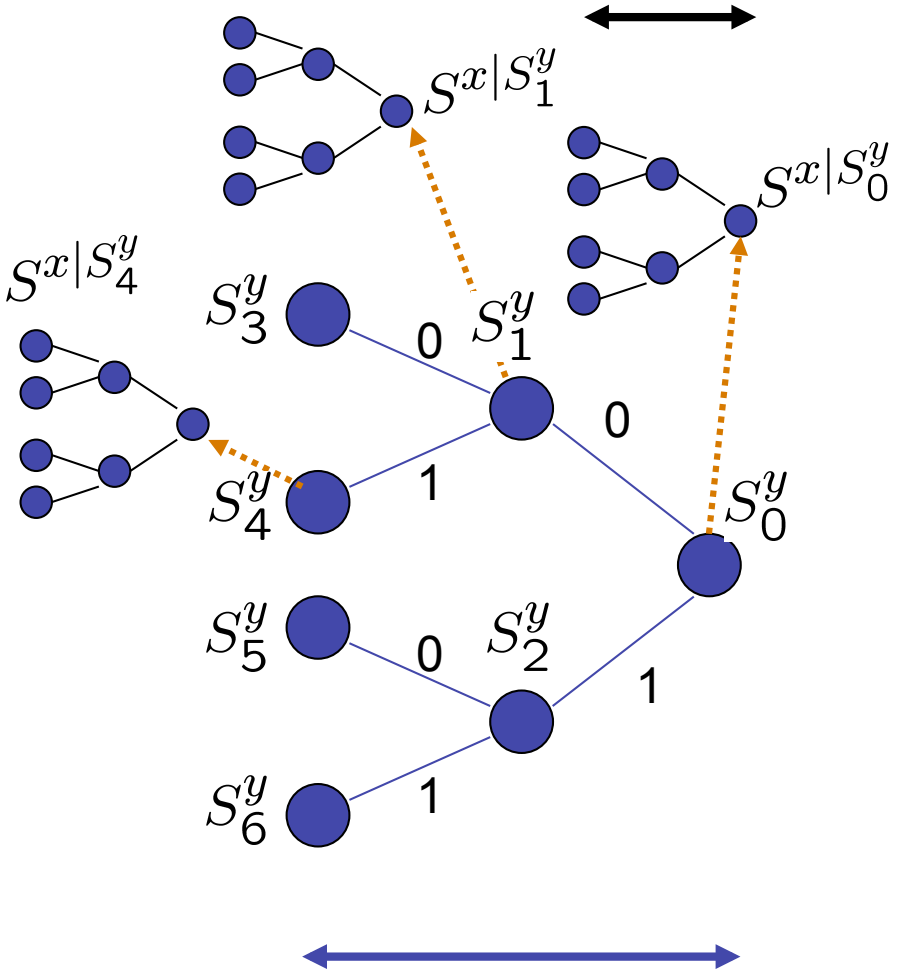
Conditional Fisher information:

$$
I_{X|Y}(\nu) = -\lim_{n\to\infty}\frac{1}{n}E_{X^nY^n}\left[\frac{\partial^2 \log P(X^n|Y^n,\nu)}{\partial\nu(\partial\nu)^T}\right]
$$

$$
E_\theta[-\log P(X^n)] = H(X^n|\theta) + \frac{k}{2}\log\frac{n}{2\pi e} + \log\frac{\sqrt{\det I_X(\theta)}}{p(\theta)} + o(1).
$$

12. Time complexity of the algorithm

d(x) : The depth of the context tree for *x*

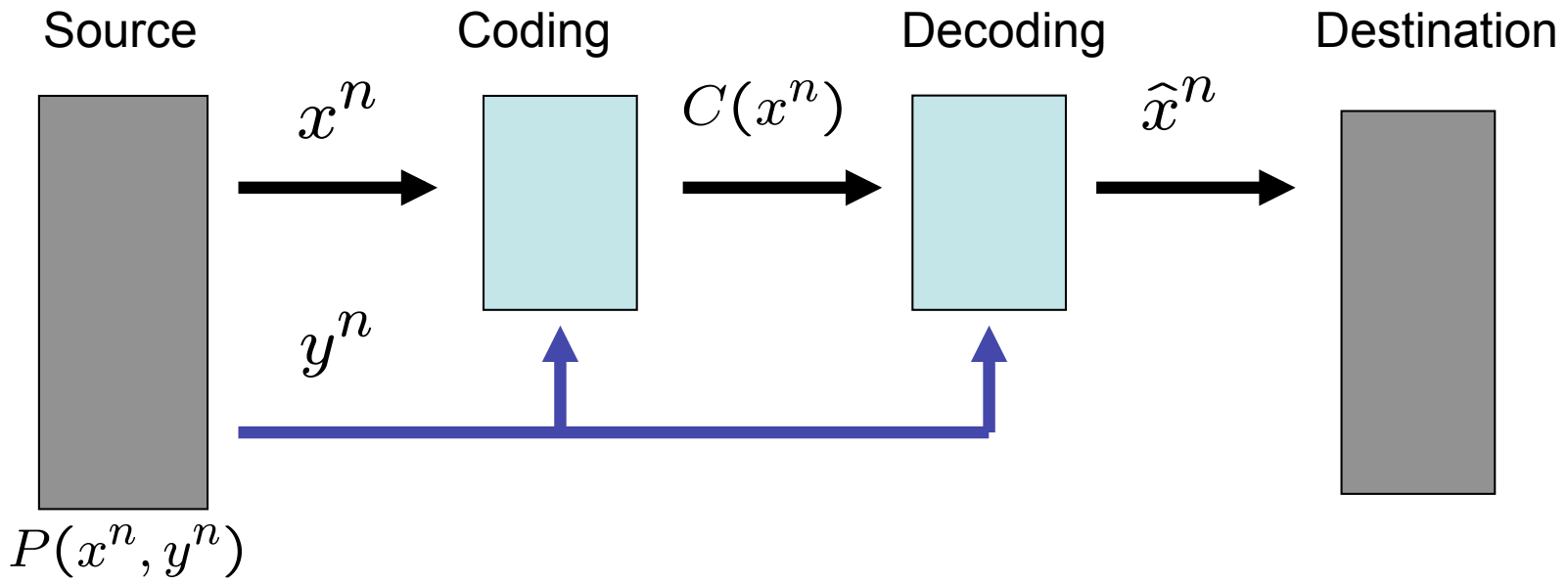


The time complexity of the algorithm: *O(d(x)d(y))*

The parallel time complexity of the algorithm: *O(d(x)+d(y))*

d(y) : The depth of the context tree for *y*

まとめ



- 補助情報付ユニバーサル符号と学習理論の関係
- 補助情報付ベイズ符号の累積log損失（符号長）の漸近評価
- 補助情報文脈木モデルとその上での効率的符号化アルゴリズム